Temporal Aggregation Bias and Inference of Causal Regulatory Networks

Stephen Bay,[1] **Lonnie Chrisman,**[1] **Andrew Pohorille,**[2] **and Jeff Shrager**[1 3]

[1]Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

[2]Center for Computational Astrobiology and Fundamental Biology
NASA Ames Research Center, M/S 239-4, Moffett Field, CA 94305

[3]Department of Plant Biology, Carnegie Institute of Washington

Abstract

Time course experiments with microarrays have begun to provide a glimpse into the dynamic behavior of gene expression. In a typical experiment, scientists use microarrays to measure the abundance of mRNA at discrete time points after the onset of a stimulus. Recently, there has been much work on using these data to infer causal regulatory networks that model how genes influence each other. However, microarray studies typically have slow sampling rates that can lead to temporal aggregation of the signal. That is, each successive sampling point represents the sum of all signal movements since the previous sample. In this paper, we show that temporal aggregation can bias algorithms for causal inference and lead them to discover spurious relations that would not be found if the signal were sampled at a faster rate. We discuss the effects of temporal aggregation on inference, the problems it creates, and potential directions for solutions.

Contact Author: Stephen Bay
Address:
Stanford University,
Computational Learning Laboratory
CSLI, Ventura Hall
Stanford, CA 94305-4115
Email: sbay@apres.stanford.edu
Phone: 650-723-1684
Fax: 650-723-0758

Temporal Aggregation Bias and Inference of Causal Regulatory Networks

Abstract

Time course experiments with microarrays have begun to provide a glimpse into the dynamic behavior of gene expression. In a typical experiment, scientists use microarrays to measure the abundance of mRNA at discrete time points after the onset of a stimulus. Recently, there has been much work on using these data to infer causal regulatory networks that model how genes influence each other. However, microarray studies typically have slow sampling rates that can lead to temporal aggregation of the signal. That is, each successive sampling point represents the sum of all signal movements since the previous sample. In this paper, we show that temporal aggregation can bias algorithms for causal inference and lead them to discover spurious relations that would not be found if the signal were sampled at a faster rate. We discuss the effects of temporal aggregation on inference, the problems it creates, and potential directions for solutions.

1 Introduction

An important step in understanding cellular functions is discovering how genes dynamically regulate their expression in response to external and internal cell signals. Experimental techniques such as microarrays have begun to provide observations of dynamic behavior by reporting expression levels on a genome-wide scale. For example, Spellman et al. [1998] and Cho et al. [1998] measured gene expression levels during the cell cycle of *S. cerevisiae*, Khodursky et al. [2000] examined how expression levels in *E. coli* change during tryptophan starved and rich conditions, and Hihara et al. [2001] examined the transient and long term response of gene expression in Cyanobacteria after exposure to high light levels.

It seems natural to use these data to uncover the regulatory dynamics of gene expression. Many researchers have proposed methods to "reverse engineer" or to learn the causal structure of the relationships between genes from time series data [D'Haeseleer *et al.*, 1999; Khan *et al.*, 2002; Murphy and Mian, 1999; Ong *et al.*, 2002; Ong and Page, 2001; van Someren *et al.*, 2000; 2001; Weaver *et al.*, 1999; Kundaje *et al.*, 2002]. Success in this task would not only allow predictions of gene expression levels under similar conditions, but also forecasting the effect of interventions such as gene deletions or forced overexpression.

One issue that is often ignored is the effect of temporal aggregation bias [Granger, 1969; Christiano and Eichenbaum, 1986; McCrorie, 2001; Gulasekaran and Abeysinghe, 2002] on the inferred regulatory structures. Temporal aggregation occurs when a process is sampled slower than the natural rate at which it is changing. For example, a hypothetical time series that could be the expression level of a gene is shown in Figure 1. Each sample, indicated by the squares, represents the net change in the signal since the last sample. In particular, the measurement at $t = 40$ is the sum or aggregate of all signal changes since $t = 30$. With samples every 10 time units, it may be possible to reconstruct the gross shape of the signal, but knowledge of the finer structure is lost. Temporal aggregation occurs in both continuous and discrete time processes when they are sampled too coarsely.

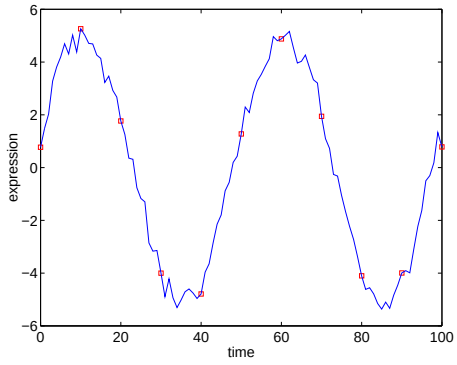


Figure 1: Slow sampling leads to temporal aggregation.

In econometrics, temporal aggregation is well known to have substantial negative effects on the inference of causal relations in discrete time models [Granger, 1969; Christiano and Eichenbaum, 1986; McCrorie, 2001; Gulasekaran and Abeysinghe, 2002]. It can lead to identification of spurious causal relations making correct inference extremely difficult. However, many papers on inferring genetic regulatory networks from temporal data have ignored the potential pitfalls of slow sampling rates. In this paper, we argue that temporal aggregation presents serious challenges for inferring dynamic regulatory networks from sampled microarray data.

In the next section, we show in principle how temporal aggregation can lead to spurious causal relationships and discuss how this affects algorithms for inferring regulatory networks. We then demonstrate on synthetic data that temporal aggregation can be a serious problem that makes inferring the causal structure of even a simple system extremely difficult. Next, we discuss the implications for learning from experimental data sets: we consider the sampling rates from current experiments and argue that temporal aggregation is a problem of practical significance under these conditions. Finally, we discuss possible approaches for dealing with temporal aggregation and suggest directions for future research.

2 Temporal Aggregation and Spurious Causality

In this section, we show how temporal aggregation can lead to incorrect inference of causal relationships between variables that have no direct link. To clarify the meaning of causal, we interpret the statement "x causes y" to mean that if the variable x were explicitly controlled then y would change in response. However, to discuss inference algorithms, we require an operational definition, and we will use the term "cause" in the sense of Granger [1969] and define it as follows.

> **Definition 1.** A time series $x(t)$ Granger causes another time series $y(t)$ if the prediction of $y(t+1)$ is improved by using present and past values of x when all other information, such as the history of y and other variables, has been considered.

We will use lower case variables, e.g., $x(t)$, to represent disaggregated signals and upper case variables like $X(T)$ to represent the aggregated signal that is recorded during sampling. The time index t corresponds to the disaggregated signal, and T corresponds to the index of the aggregated signal.

Temporal aggregation can lead algorithms to infer spurious Granger causality relationships. That is, in an aggregated signal it may appear that $X(T)$ causes $Z(T)$ because $X(T)$ improves the prediction of $Z(T+1)$ even when all other information has been considered. However, if the data were analyzed at a finer sampling resolution, i.e., in the disaggregated form, then $x(t)$ would not improve the prediction of $z(t+1)$ and thus would not be considered a cause.

Granger causality is directly related to notions of causality in most frameworks used for analyzing gene expression data. For example, consider rough network models [van Someren *et al.*, 2001], which represent a wide class of models with linear relationships between genes (or other chemical species). These models have the form [1]

$$X_i(T+1) = g(\sum_{j=1}^{J} W_{ij} X_j(T)) \,, \qquad (1)$$

where a non-zero W_{ij} means that gene X_j is a direct cause of gene X_i. The function $g(\cdot)$ can be any monotonic activation function such as a sigmoid. The parameters and structure of these models are typically learned by minimizing a score function based on prediction error, which often incorporates a penalty for complexity. Thus during model search, a hypothesized causal relation must improve prediction to be included as a direct link (i.e., a non-zero W_{ij}).

Another common framework for inferring regulatory networks is based on the causal interpretation of dynamic Bayesian networks (DBN) [Murphy and Mian, 1999; Ong and Page, 2001; Ong *et al.*, 2002; Khan *et al.*, 2002] learned from data. A Bayesian network [Pearl, 1988] is a graphical model in which nodes represent variables and the pattern of directed links represent conditional independence relations. A dynamic Bayesian network extends the Bayesian network formalism to model how variables evolve over time. DBNs can be represented as a directed graph in which each variable is represented by a node for every time point. When interpreted causally, a link from variable $X(t=i)$ to $Y(t=j)$ indicates that X at time i is a direct cause of Y at time j. Each variable has a local model that determines its value as a probabilistic function of its causes (parents).

There are two main approaches to learning the structure of a Bayesian network from data. The first is a score based approach (e.g., [Cooper and Herskovits, 1992; Heckerman, 1999; Friedman *et al.*, 2000]) where the goal is to find the network that yields the best predictions of the data. This ties into Granger causality in the same way as rough network models: i.e., a hypothesized causal relation must improve prediction when all other information has been considered. The second method of learning is based on matching the conditional independence relations observed between variables in the data with those entailed by the network structure (e.g., [Glymour *et al.*, 1987; Saavedra *et al.*, 2001]). Saavedra et al. [2001] took this approach to infer regulatory relations from yeast cell cycle data. As with scoring approaches, this also ties to Granger causality because testing for conditional independence is equivalent to testing if a variable has predictive power. Under a linear model, two variables are conditionally independent if and only if they have a zero partial correlation [Pearl, 1998].

A useful way of understanding why spurious causal relations occur under temporal aggregation is to view these systems graphically in the same manner as a dynamic Bayesian network. Many formalisms for inferring genetic regulatory networks can be considered special cases of Dynamic Bayes networks [Murphy and Mian, 1999] including Boolean networks [Liang *et al.*, 1998; Akutsu *et al.*, 1999; Somogyi and Sniegoski, 1996], rough network models, both linear [D'Haeseleer *et al.*, 1999] and non-linear models [Weaver *et al.*, 1999], as well as the vector autoregressive models commonly used to analyze economic time series.

In Bayesian and dynamic Bayesian networks one can view the graphical representation and directly determine conditional independence relationships and hence whether a variable is predictive of another. Two variables A and B are conditionally independent if there is an intermediate variable C on all undirected paths between them such that

1. C does not have converging arrows (i.e., $\leftarrow C \leftarrow$, $\rightarrow C \rightarrow$, $\leftarrow C \rightarrow$) and C is observed;

[1] This is slight a simplification of the equations presented by van Someren et al. [2001].

2. C has converging arrows ($\rightarrow C \leftarrow$) and neither C nor its descendents are observed.

This criterion is known as *d-separation* [Pearl, 1998].

Figure 2a shows a dynamic Bayes net representation of a three variable system: $x(t) = f(x(t–1))$, $y = f(x(t–1), y(t–1))$, and $z = f(y(t-1), z(t-1))$. For each time step, the network has been "unrolled" and has a node representing the value of each variable at that time point. Arrows represent functional dependencies which may be across time steps. For example, the value of $y(t=2)$ is a function of $x(t=1)$ and $y(t=1)$. In this figure, we have observations at every time step. The variable z is conditionally independent of x given y (i.e., all paths from x to z are blocked by y under condition 1). Hence, analyzing the data generated by this process would let us infer that x does not Granger cause z because they are conditionally independent given y (x adds no predictive power once y is known) and that y Granger causes z.

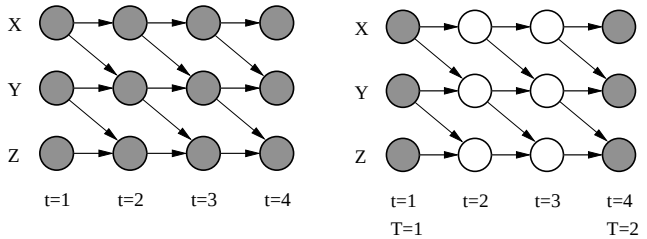


Figure 2: Dynamic Bayes net representation of a three variable system. Observed nodes are shaded; unobserved nodes are clear. (a) Sampling at every time step. (b) Sampling with temporal aggregation.

When the signal is aggregated, we miss observations and we might have the situation in Figure 2b where the observations at $t = 2$ and $t = 3$ are missing. With aggregation, $X(T=1)$ is no longer conditionally independent of $Z(T=2)$ given the history of Y. There are several unblocked paths from X to Z such as $X(T=1)$, $y(t=2)$, $z(t=3)$, $Z(T=2)$. Thus, it will appear that X Granger causes Z even though this is not true in the disaggregated case. Essentially, $X(T=1)$ gives information about the missed samples at $t=2$ and $t=3$. This in turn provides information about the value of Z in the next sample in the aggregated data at $T = 2$. If the latent unobserved steps are not modeled it will appear as if X is a direct cause of Z.

Aggregation will lead to correlation between X and Z that cannot be explained by Y. In this case, algorithms based on predictive scoring or conditional independence relationships will infer a direct, but spurious, link between X and Z. In general, with any set of variables that are individually correlated, aggregation will lead them to look like causes of each other [McCrorie, 2001].

In bivariate systems with one-directional causality, aggregation can actually lead to the inference of a spurious bidirectional causality relation (a feedback cycle). For example, consider Figure 3a. Here x is a cause of y as $x(t)$ provides useful information for predicting $y(t+1)$ and therefore x Granger causes y. However, $y(t)$ does not produce information useful for predicting the future values of x, so y does not Granger cause x. In Figure 3b, the aggregated variable X still provides information about future values of Y, so as before X is a cause of Y. However, now that some observations are missing, Y can provide useful information for predicting future values of X through many paths such as $Y(T=1)$, $x(t=1)$, $x(t=3)$, $X(T=2)$. Essentially, the observed values of Y give information about the missed samples of X which then helps in predicting future values. Therefore, when analyzing the aggregated data, Y will Granger cause X. Since it appears that X causes Y, and Y causes X, we would incorrectly infer bidirectional causality (a feedback cycle).

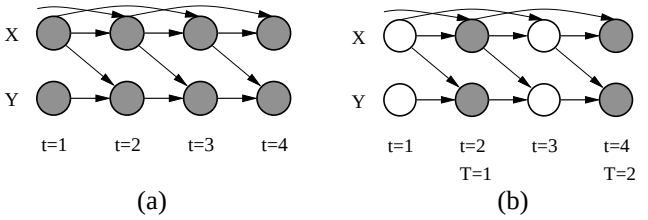


Figure 3: Dynamic Bayes net representation of a two variable system. Observed nodes are shaded; unobserved nodes are clear. (a) Sampling at every time step. (b) Sampling with temporal aggregation.

In summary, aggregation causes variables not directly related to each other to have predictive power that cannot be explained away by other variables. This leads most algorithms to infer a direct, but spurious, link between them.

3 Experiments with Synthetic Data

In this section, we describe experiments with a simple synthetic system and we show that aggregation can lead to spurious inferences about causal relationships. We begin by considering the discrete-time three variable system:

$$x(t) = Asin(Bt) + \epsilon_1(t) \quad (2)$$
$$y(t) = x(t-d_1) + \epsilon_2(t) \quad (3)$$
$$z(t) = y(t-d_2) + \epsilon_3(t) \quad (4)$$

We interpret this set of equations in the same manner as structural equation models [Bollen, 1989; Pearl, 1998] where the variable on the left hand side of the equality is caused by the variables on the right hand side. The variable x is a sinusoid with amplitude A, period $2\pi/B$; x causes y but the effect is delayed by time d_1; and y causes z with a delay of d_2. The noise variables $\epsilon_1(t)$, $\epsilon_2(t)$, and $\epsilon_3(t)$ represent errors caused by omitted factors and are uncorrelated. The variable x is not a cause of z and only affects z indirectly through y.

We will use this system to generate synthetic data and show that spurious Granger causality relationships can occur with different methods for causal learning. Specifically, we will examine how methods for learning vector autoregressive models and methods based on partial correlation behave on aggregated data.

For our experiments, we used Equations 2–4 to generate data and we systematically varied the aggregation rate by sampling every 1,2,...,10 samples of the original time index.

We set the random noise variables $\epsilon_i(t)$ to $N(0, \sigma^2)$ where $\sigma^2 = \{0.01, 0.25, 1, 4\}$. We randomly set the delays d_1 and d_2 to an integer between 1 and 20. We set the amplitude, A, of the sine to 5 and the period to 40. In Figure 1, we show an example of this signal for $\sigma^2 = 0.25$. Note that we do not claim that these synthetic data accurately describe real biological systems, but clearly a real biological system will have a more complicated structure making the causal learning problem even harder.

3.1 Vector Autoregressive Models

In our first experiment, we try to learn a vector autoregressive model for Z with the following form,

$$Z(T) = w + \sum_{i=1}^{p} \alpha_i X(T-i) + \sum_{i=1}^{p} \beta_i Y(T-i) + \sum_{i=1}^{p} \gamma_i Z(T-i) \quad (5)$$

That is, the value of the variable Z at time T is a linear combination of past values of the variables X, Y, and Z plus a constant w. The order of the system, p, refers to the maximum delay in the aggregated time index T. For disaggregated data, this model can exactly represent the generating process in Equations 2-4. This model can be represented as a dynamic Bayes network, and if the order of the system is limited to 1, a rough network model.

In any model inferred from the data, all of the α_i coefficients should be zero since $x(t-d)$ does not directly affect $z(t)$ for any value of d. We generated 100 data sets for each combination of aggregation rate and noise level according to the procedure outlined earlier. Each data set was 500 points long in the aggregated time index T and we ignored the beginning points to eliminate problems with the initial values of y and z. For each data set, we fit an vector autoregressive model for Z with the Matlab package ARfit [Schneider and Neumaier, 2001]. ARfit uses a stepwise least squares procedure to estimate the parameters w, α_i, β_i, and γ_i.

In Figure 4, we show the number of times ARfit inferred that X was a cause of Z for a given aggregation rate with $\sigma^2 = 0.25$. The results for other noise levels were similar. We considered α_i to represent a causal influence if its magnitude was significantly different from zero (i.e., the three standard deviation confidence intervals[2] did not include zero).

With no aggregation (aggregation rate = 1), ARfit correctly generates a model with only Y as a predictor. However, even a slight amount of aggregation causes an incorrect causal inference that X is directly related to Z. Note that the spurious causal inference is not explained by variance in parameter estimates and the aggregation problem cannot be solved by simply obtaining larger amounts of data.

In addition to creating spurious causal links between variables, aggregation also leads to known causal interactions disappearing. In Figure 5, we plotted the frequency of Y appearing as a cause of Z, i.e., $\beta_i \neq 0$ for at least one i. With no aggregation Y almost always appears as a cause. With aggregation (rate ≥ 2), Y frequently failed to appear as a cause.

What this means for causal learning is that when the data signal is aggregated, variables not directly linked can have very strong predictive power for each other and this cannot be explained away with mediating factors. This leads to spurious links under predictive score based measures.

[2] uncorrected for multiple hypotheses

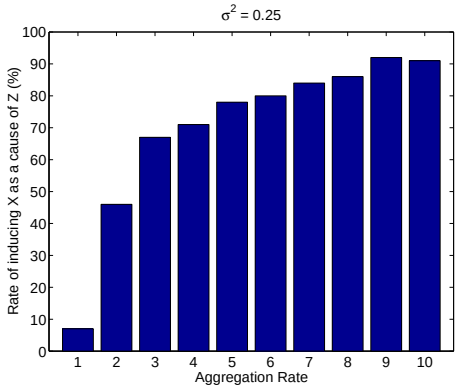


Figure 4: Rate of inferring X as a (spurious) cause of Z.

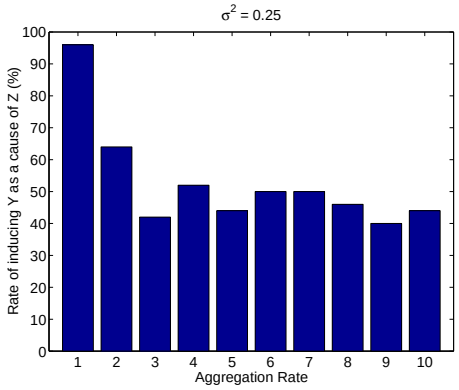


Figure 5: Rate of inferring Y as a cause of Z.

3.2 Partial Correlation

Another common method of causal learning is to use correlations and partial correlations to tease out the causal relations between variables. Correlation measures the predictive power of two variables for each other under a linear model. Partial correlation measures predictive power of two variables accounting for the effects of other variables [Anderson, 1984]. Partial correlation analysis is widely used to determine causal effects from data (e.g., [Glymour *et al.*, 1987; Dahlhaus, 2000]) and for model selection in autoregressive processes [Barndorff-Nielsen and Schou, 1973]. If a correlation between two variables x and z is explained away by a third variable y (i.e, the partial correlation $\rho_{xz.y} = 0$), then x and z are not directly related. With time series data, two variables should be uncorrelated for all possible time shifts once the effects of other variables are removed.

For the second experiment, we examined how the partial correlation of variables X and Z controlling for Y varied with aggregation. As before, we generated 100 data sets for each combination of aggregation rate and noise level. For each data set we measured the sample partial correlation of X and Z controlling for Y as follows: first, we found the lag, d_z, between X and Z that maximizes their cross-correlation

and then found the minimum partial correlation controlling for Y with a lag $[0, d_z]$. In Figure 6 we show the minimum partial correlations averaged across the 100 data sets with $\sigma^2 = 0.25$. The results for other noise values were similar.

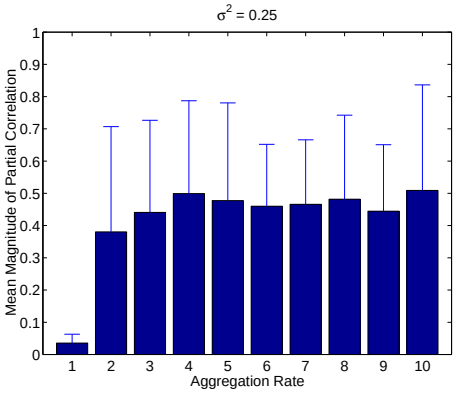


Figure 6: Mean magnitude of the partial correlation of X and Z controlling for Y. The vertical lines represent one standard deviation of the partial correlation coefficient.

The results show that Y only explains the apparent correlation between X and Z when there is no aggregation (aggregation rate = 1). As soon as the signal is aggregated, the partial correlation rises to substantial levels. This indicates an interaction between X and Z which cannot be explained by Y thus leading to a spurious causal link.

What this means for causal learning is that temporal aggregation destroys conditional independence relations in the data and thus indirectly related variables will have a dependency which cannot be explained away by other variables. This leads to incorrect inference of a direct causal link.

3.3 Interpolation

In the previous two experiments, we showed that in the aggregated time index autoregressive and partial correlation methods would frequently find spurious causal relations. Another alternative is developing models in the original time index by replacing the missed samples with interpolated values. We repeated the experiments by interpolating the aggregated data with cubic splines to reconstruct the missing data samples.

With autoregressive models, interpolation did not help at low noise values ($\sigma^2 = 0.01$), and X was inferred as a spurious cause of Z in the majority of trials. At higher noise values ($\sigma^2 = 4$), X was less likely to be inferred as a cause (compared with no interpolation), although still substantially more frequently than if there were no aggregation.

One significant change with interpolation is that past values of Z were always inferred as causes of $Z(t)$. That is, in every trial with an aggregation rate greater than or equal to 2, $Z(t-d)$ for at least one value of d was inferred as a cause of $Z(t)$ (with no aggregation $Z(t-d)$ is almost never inferred as a cause). Clearly, this is an incorrect inference as $Z(t)$ has no dependence on historical values of Z in the original equations. The spurious dependence arises because interpolation by definition is the act of estimating the value of a variable between observed points (i.e., past and future values). The autoregressive model is recovering portions of the mathematical structure applied to interpolate the data.

We repeated the partial correlation experiment with interpolated data. Again, interpolation did not reduce the aggregation problem and the results were similar to those in Figure 6.

To summarize, interpolation does not mitigate the aggregation problem for inference with autoregressive models or partial correlation. Interpolation however, introduces additional relationships between the past, present, and future values of a variable thus further confounding causal reconstruction.

4 Temporal Aggregation in Gene Expression Data

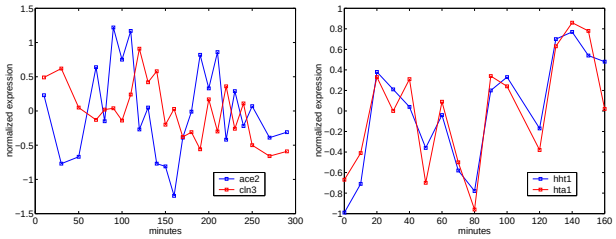
We have shown that temporal aggregation can make causal inference extremely difficult as it is easy to infer spurious relationships not present in the disaggregated data. In this section, we argue that temporal aggregation is a serious problem for inferring causal regulatory networks from microarray expression data. The central question is how fast are expression levels changing and how does this compare with the sampling rate? We argue that present methods are insufficient to reconstruct causal regulatory networks at current sampling rates in the absence of other knowledge or experimental designs.

4.1 Sampling Rates and Gene Expression Levels

Time course experiments that sample the state of gene expression with microarrays have slow sampling rates. Studies such as those conducted by Spellman et al. [1998] and Cho et al. [1998] sampled a system every 7 to 30 minutes. Khodursky et al. [2000] used uneven sampling intervals and took measurements at 5, 15, 30, and 60 minutes after their experiment's start. Likewise, Hihara et al. [2001] used uneven samples at 15 min., 60 min., 6 hours, and 15 hours.

The data from these experiments indicate that current sampling rates cannot fully capture the changes of mRNA levels and at best allow only coarse reconstruction of the original signal. For example, Spellman et al. [1998] and Cho et al. [1998] measured gene expression of *S. cerevisiae* as it progressed through the cell cycle. In total, they conducted four experiments, each corresponding to a different method of synchronizing the cells: α factor, cdc15, cdc28, and elutriation. The sampling varied from every 7 minutes (α factor) to every 30 minutes (elutriation) and covered from 14 to 24 time points. These are among the fastest published sampling rates for microarray studies. In Figure 7, we plot the expression of several genes. In part (a), we show ace2, a transcription factor, and its target cln3 [Simon *et al.*, 2001]. In part (b), we show hta1 and hht1 which code for histones related to chromatin structure that are upregulated in the S phase of the cell cycle [Simon *et al.*, 2001].

Clearly, the signal varies rapidly between adjacent time points and does not instill confidence that the expression levels are adequately sampled. Although the large variation could be caused by measurement error, or a common noise source unrelated to the signal, such as a bad chip, it is also likely that rapid changes in expression levels are not fully captured and the signal is aggregated between time points.



(a) (b)

Figure 7: Expression of genes during the cell cycle of *S. Cerevisiae*. (a) ace2 and cln3 under cdc15 synchronization. (b) hht1 and hta1 under cdc28 synchronization.

Finally, the rate of gene expression can respond abruptly to changes in cell conditions. For example, Lu et al. [2002] conducted studies with green fluorescent protein (GFP), a reporter gene product that can be observed visually and thus making it possible to obtain near real time and continuous measurements of gene expression. They used GFP to measure the activity of the araBAD operon in E. coli. An operon is set of contiguous genes that are transcribed together. Lu et al. found that the rate of expression could abruptly increase in response to arabinose and decrease in response to glucose. Any algorithm that attempts to determine causal structure would need data that detect these change points.

4.2 Effects on Reconstruction of Causal Relations

Temporal aggregation may also explain the poor results in causal learning reported by computational approaches tested on real microarray data. Specifically, Ong and Page [2001], and in an expanded study Ong, Glasner, and Page [2002], evaluated the effectiveness of dynamic Bayesian networks for inferring the regulatory structure of tryptophan metabolism in *E. Coli* with sparse time course data and found many apparent spurious causal relations.

In their study, they used a DBN to learn how the operons in *E. Coli* regulate each other. For example, the trp operon would include the genes trpA, trpB, trpC, trpD, and trpE. In their network, Ong and Page included variables that represented the activity level of 141 known operons determined from previous work [Salgado *et al.*, 1999; Craven *et al.*, 2000] and the genes in each operon. The level of expression of each gene is controlled by the activity of its operon. Each operon, in turn, is controlled by its activity level at the previous time step and possibly by one other operon at the previous time step. Note that the activity levels of the operons are not directly observable, but are inferred from the expression levels of the controlled genes.

To learn causal structure, Ong and Page used data from Khodursky et al. [2000] who examined how gene expression levels in *E. coli* change during tryptophan starved and rich conditions. Khodursky exposed *E. coli* to varying conditions and measured gene expression levels at 4 time points: 5, 15, 30, and 60 minutes after the experiment's start. In their first study Ong and Page [2001] used two time series of 4 data points each for tryptophan rich and starved conditions. In the second study, Ong, Glasner, and Page [2002] had an additional 4 points for tryptophan starved conditions.

They compared their discovered network with the biological literature (summarized in [Khodursky *et al.*, 2000]) and found that "in every case above where a tryptophan-related operon was chosen as the best or second-best parent of another tryptophan related operon, the relationship between the operons in the regulatory pathway either flows in the opposite direction or is a relationship of indirect influence rather than direct"[Ong and Page, 2001]. For example, the DBN learning algorithms inferred that the trp operon was a probable direct parent of trpR, which is a reversal of the known regulatory relation. It also inferred that several known siblings were regulators of each other: the trp operon was found to be a parent of mtr although both are believed to be regulated by trpR.

These results are consistent with temporal aggregation, however other factors such as the limited number of parents for each operon or possible errors in assigning genes to operons could also contribute to the spurious results.

5 Implications for Future Research

We believe microarray data alone are insufficient for causal inference of time dynamics. The currently used sampling rates appear to be too low (e.g., once every 10-30 minutes) to distinguish causality from correlation given how fast expression can change. At a minimum then, researchers proposing new algorithms should investigate their robustness and sensitivity to temporal aggregation of the data.

We envision three venues for addressing the problem of temporal aggregation: (1) Improving the sampling rates with alternative measurement technologies. (2) Incorporating data from experimental designs with causal interventions. (3) Using background knowledge in conjunction with the data signal.

5.1 Improving Sampling Rates

Improving sampling rates to adequately capture the expression signal will eliminate aggregation bias, but this may be difficult to achieve with current experimental procedures for microarrays. The primary limitation is cost, for both the microarrays which may be several hundred dollars each, and the experimental apparatus to support the biological materials needed for each data point. However, there are other techniques, such as inserting reporter genes coding for GFP [Lu *et al.*, 2002], that allow continuous and *in vivo* measurements of gene expression. These methods appear very promising, but they have some drawbacks for genome-wide analysis. For example, GFP is usually inserted into cells on a plasmid that tracks a specific promoter and not an individual gene. Thus, GFP will not capture modifications that change the rate of mRNA expression for the genes within an operon. It may also be difficult to get simultaneous measurements from multiple promoters thus making system-wide analysis impossible. However, one could measure a subset of genes that relate to a specific system of interest. For example, Kalir et al. [2001] used multiple GFP reporter plasmids with different promoters to obtain measurements of 14 operons involved with the synthesis of flagella in *E. coli*.

5.2 Experimental Designs with Causal Interventions

Temporal aggregation makes inference difficult because the aggregated signal does not adequately capture fluctuations from noise terms that are essential for explaining away the predictive power of indirectly related variables. However, another promising approach is to use data from experimental designs with causal interventions which may not require fine sampling to determine causal effects. For example, Kalir et al. [2001] used timing studies after a perturbation to partially elucidate the causal ordering of gene regulation in the construction of flagella. Stationary phase *E.coli* were placed into a fresh medium which stimulated growth and Kalir et al. used statistics from the rise time of fluorescence to determine the ordering of promoters in gene regulation. Although this did not completely tease apart the causal structure of regulation, it did provide a partial order on the regulation that can occur. Vance, Arkin, and Ross [2002] took a similar approach to determine the causal connectivity of reaction networks from pulses in the concentration of chemical species.

5.3 Using Background Knowledge

Using background knowledge in conjunction with the data signal can reduce the chance of inferring spurious causal relationships by constraining inference algorithms. For example, Ong, Glasner, and Page [2002] in their experiment with regulation in *E. coli* (described Section 4.2) used the operon structure to eliminate " 'useless' arcs between genes in the same operon". Genes within an operon are typically highly correlated as they are controlled with a common promoter, but they do not generally regulate each other. Inference from aggregated time series measurements would then likely result in many spurious links between these genes, but enforcing the constraint that genes in the same operon cannot regulate each other eliminates this problem.

Another situation where background knowledge may help reduce spurious inferences is when a causal ordering on variables is known. For instance, in *E. coli* the order in which genes are expressed to construct flagella is partially known [Kalir *et al.*, 2001]. If it is known that Y comes before Z, we do not need to test if Z is a possible parent of Y. This prevents us from inferring that Z is a cause of Y even if the aggregated data indicates that this is the case.

6 Conclusions

In this paper, we discussed temporal aggregation which occurs when a signal is sampled too coarsely, and how it can bias causal inference. We showed in principle that aggregation could lead to unexplainable correlations between indirectly connected variables leading algorithms to infer a direct but spurious causal relationship. We demonstrated with a simple synthetic system, that even when the data is generated from a model within the representational class that the inference algorithm is designed to recover, aggregation can cause spurious causal relations. Interpolation does not mitigate this problem and can make it even worse by introducing additional correlations that do not exist in the true system. In future work on inferring regulatory networks, we believe that it will be necessary to design algorithms that recognize the limitations arising from temporal aggregation and work within these constraints.

Acknowledgments

This work was supported by the NASA Biomolecular Systems Research Program.

References

[Akutsu *et al.*, 1999] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, pages 17–28, 1999.

[Anderson, 1984] T. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, NY, 1984.

[Barndorff-Nielsen and Schou, 1973] O. Barndorff-Nielsen and G. Schou. On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3:408–419, 1973.

[Bollen, 1989] K. Bollen. *Structural Equations with Latent Variables*. Wiley, 1989.

[Cho *et al.*, 1998] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[Christiano and Eichenbaum, 1986] L. J. Christiano and M. Eichenbaum. Temporal aggregation and structural inference in macroeconomics. Technical Report 60, National Bureau of Economic Research, 1986.

[Cooper and Herskovits, 1992] G. F. Cooper and E. H. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[Craven *et al.*, 2000] M. Craven, D. Page, J. Shavlik, J. Bockhorst, and J. Glasner. A probabilistic learning approach to whole-genome operon prediction. In *Proc. of the 8th Int. Conf. on Intelligent Systems for Molecular Biology*, pages 116–127, 2000.

[Dahlhaus, 2000] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172, 2000.

[D'Haeseleer *et al.*, 1999] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels durning CNS development and injury. In *Pacific Symposium on Biocomputing*, pages 41–52, 1999.

[Friedman *et al.*, 2000] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3/4):601–620, 2000.

[Glymour *et al.*, 1987] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, 1987.

[Granger, 1969] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[Gulasekaran and Abeysinghe, 2002] R. Gulasekaran and T. Abeysinghe. The distortionary effects of temporal aggregation on granger causality. Technical Report 0204, Department of Economics, National University of Singapore, 2002.

[Heckerman, 1999] D. Heckerman. A tutorial on learning Bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1999.

[Hihara *et al.*, 2001] Y. Hihara, A. Kamei, M. Kanehisa, A. Kaplan, and M. Ikeuchi. DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. *The Plant Cell*, 13:793–806, 2001.

[Kalir *et al.*, 2001] S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292:2080–2083, 2001.

[Khan *et al.*, 2002] R. Khan, Y. Zeng, J. Garcia-Frias, and G. Gao. A Bayesian modeling framework for genetic regulation. In *Proc. of the IEEE Computer Society Bioinformatics Conf.*, 2002.

[Khodursky *et al.*, 2000] A. Khodursky, B. J. Peter, N. R. Cozzarelli, D. Botstein, P. O. Brown, and C. Yanofsky. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. *Proc. of the National Academy of Science*, 97(22):12170–12175, 2000.

[Kundaje *et al.*, 2002] A. Kundaje, O. Antar, T. Jebara, and C. Leslie. Learning regulatory networks from sparsely sampled time series expression data. Technical report, 2002.

[Liang *et al.*, 1998] S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of geneticnetwork architectures. In *Pacific Symposium on Biocomputing*, volume 3, pages 18–29, 1998.

[Lu *et al.*, 2002] C. Lu, C. R. Albano, W. E. Bentley, and G. Rao. Differential rates of gene expression monitored by green flourescent protein. *Biotechnology and Bioengineering*, 79(4):429–437, 2002.

[McCrorie, 2001] J. R. McCrorie. Granger causality and the sampling of economic processes. In *Proc. of the 12th Conf. on Causality and Exogeneity in Econometrics*, 2001.

[Murphy and Mian, 1999] K. Murphy and S. Mian. Modelling gene expression data using dynamic Bayesian networks. Technical report, University of California, Berkeley, 1999.

[Ong and Page, 2001] I. M. Ong and D. Page. Inferring regulatory pathways in E. coli using dynamic Bayesian networks. Technical Report 1426, University of Wisconsin-Madison, May 2001.

[Ong *et al.*, 2002] I. M. Ong, J. D. Glasner, and D. Page. Modeling regulatory pathways in E. coli from time series expression profiles. In *Proc. of the 10th Int. Conf. on Intelligent Systems for Molecular Biology*, 2002.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.

[Pearl, 1998] J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284, 1998.

[Saavedra *et al.*, 2001] R. Saavedra, P. Spirtes, R. Ramsey, and C. Glymour. Issues in learning gene regulation from microarray databases. Technical Report IHMC-TR-030101-01, Institute for Human and Machine Cognition, 2001.

[Salgado *et al.*, 1999] H. Salgado, A. Santos, U. Garza-Ramos, J. van Helden, E. Diaz, and J. Collado-Vides. Regulondb (version 2.0): a database on transcriptional regulation in Escherichia coli. *Proc. of the National Academy of Science*, 27:59–60, 1999.

[Schneider and Neumaier, 2001] T. Schneider and A. Neumaier. Algorithm 808: ARFIT – a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software*, 27(1):58–65, 2001.

[Simon *et al.*, 2001] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

[Somogyi and Sniegoski, 1996] R. Somogyi and C. A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, 1996.

[Spellman *et al.*, 1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[van Someren *et al.*, 2000] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders. Linear modeling of genetic networks from experimental data. In *Proc. of the 8th Int. Conf. on Intelligent Systems for Molecular Biology*, pages 355–366, 2000.

[van Someren *et al.*, 2001] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders. Genetic network models: A comparative study. In *Proc. of SPIE, Micro-arrays: Optical Technologies and Informatics (BIOS01)*, volume 4266, pages 236–247, 2001.

[Vance *et al.*, 2002] W. Vance, A. Arkin, and J. Ross. Determination of causal connectivities of species in reaction networks. *Proc. of the National Academy of Sciences*, 99(9):5816–5821, 2002.

[Weaver *et al.*, 1999] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. In *Pacific Symposium on Biocomputing*, pages 112–123, 1999.